ExaBayes: Massively Parallel Bayesian Tree Inference for the Whole-Genome Era

Andre J. Aberer,*[,1] Kassian Kobert,[1] and Alexandros Stamatakis[1,2]

[1]Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

[2]Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany

***Corresponding author:** E-mail: andre.aberer@h-its.org.

Associate editor: Xun Gu

Abstract

Modern sequencing technology now allows biologists to collect the entirety of molecular evidence for reconstructing evolutionary trees. We introduce a novel, user-friendly software package engineered for conducting state-of-the-art Bayesian tree inferences on data sets of arbitrary size. Our software introduces a nonblocking parallelization of Metropolis-coupled chains, modifications for efficient analyses of data sets comprising thousands of partitions and memory saving techniques. We report on first experiences with Bayesian inferences at the whole-genome level using the SuperMUC supercomputer and simulated data.

Key words: **software, Bayesian statistics, phylogenetic inference, whole-genome analyses, parallelization.**

The task of resolving the tree of life of extant species remains one of the grand challenges in evolutionary biology. As the number of trees grows superexponentially with the number of species for which an evolutionary tree is reconstructed, tree inference is considered a hard problem in computer science. The plethora of algorithmic challenges associated with phylogenetic trees and their efficient computation gave rise to the discipline of "phyloinformatics." Likelihood-based statistical methods are highly popular because they can incorporate complex evolutionary models. Popular likelihood-based tools comprise methods for maximum-likelihood (ML) estimation, such as `RAxML` (Stamatakis 2014) and `PhyML` (Guindon et al. 2010), as well as Bayesian inference packages, such as `MrBayes` (Ronquist et al. 2012) and `BEAST` (Drummond et al. 2012). Likelihood-based methods can, in fact, unravel the true evolutionary tree given enough data and the appropriate model (Yang 1994b). Although the ML approach strives to optimize the likelihood of the tree and model given the data, Bayesian inference uses Markov chain Monte Carlo (MCMC) (Hastings 1970) sampling to integrate over the entire parameter space (e.g., model parameters and tree topologies) given the data and subjective prior assumptions.

Inexpensive wet-lab sequencing technologies allow amassing molecular evidence from hundreds of genes (DellAmpio et al. 2014). This amount of data is often required for resolving ancient radiations (Dunn et al. 2008). Hence, so-called phylogenomic data sets are being increasingly used to disentangle evolution. Collaborative efforts such as the 1K Insect Transcriptome Evolution (http://1kite.org, last accessed August 13, 2014) project aim at assembling the entirety of molecular sequence data to infer accurate phylogenies. Conducting such phylogenomic analyses is challenging due to exorbitant runtime and memory requirements. In a recent study (Rinke et al. 2013) with more than 2,000 microbial species and more than 8,000 amino acid alignment characters, the authors did not complement ML trees by Bayesian tree inferences. The reason for this was the limited ability of current Bayesian inference software to leverage supercomputer resources (often several CPU-months or CPU-years are required) and to accommodate the excessive main memory requirements which for this example are in the order of 10–20 GB.

Novel Approaches and Software Features

We resolve these computational limitations of Bayesian inference tools by introducing `ExaBayes`, a software package engineered for efficient Bayesian tree inference on data sets of almost arbitrary size. `ExaBayes` can conduct Bayesian analyses for the most widely used priors, models, and input data types. These comprise Dirichlet, exponential and uniform priors, the general time-reversible (GTR) model of nucleotide substitution (Tavaré 1986), the Γ model of rate heterogeneity (Yang 1994a), and unconstrained branch length sampling. For protein data, we offer 18 commonly used fixed-rate substitution matrices as well as a GTR model. We also implemented a comprehensive set of topological proposals that assure rapid convergence and adapted these for massively parallel execution.

`ExaBayes` is a self-contained software package, that is, it comprises several postprocessing methods such as standalone tools for building consensus trees, for assessing topological convergence among independent runs (Lakner et al. 2008), and for extracting sample statistics. The output format is compatible with popular visualization tools, such as `FigTree` (Rambaut 2014) or `Tracer` (Rambaut and Drummond 2007).

Runtime as well as memory requirements of Bayesian inference is largely dominated by evaluating the

Letter

Fast Track

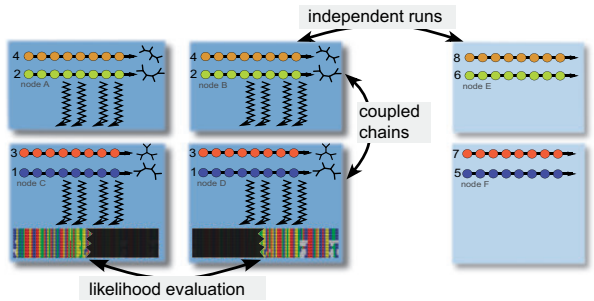

FIG. 1. The three layers of parallelism employed by ExaBayes (distributed likelihood evaluation, distributed Metropolis-coupled chains, and distributed independent analyses).

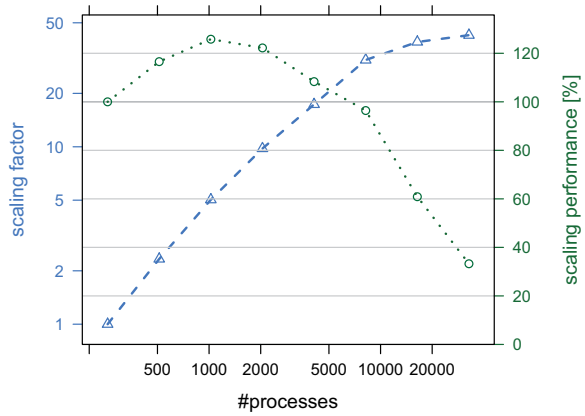

FIG. 2. Scaling factor (sequential runtime divided by parallel runtime) and efficiency (scaling factor divided by number of processes) for executing ExaBayes on 256 cores up to 32,768 cores on a 200 species alignment with 500,000 characters.

phylogenetic-likelihood function. In ExaBayes, we deploy the efficient-likelihood kernel developed for RAxML (Stamatakis 2014) that allows fully leveraging the computational power of SSE and AVX vector units on modern CPUs. The sequential AVX version of ExaBayes outperforms MrBayes for DNA data, whereas for protein data MrBayes is faster (see Supplementary Material online, for detailed discussions on this and further results). The central contribution of ExaBayes lies in its scalability.

ExaBayes implements three layers of parallelism (see fig. 1). For instance, the parallel version of ExaBayes allows to distribute the likelihood calculations across several processors and across several computing nodes in a cluster. Thus, the resources of an entire cluster or supercomputer can be used to accommodate the runtime and memory requirements of large-scale alignments. Parallel execution reduces runtimes almost linearly to the number of CPUs used. For a single chain and on a simulated data set with 200 species and 500,000 DNA characters, using the parallel version of ExaBayes reduces runtimes from 1 day and 4 h on one core to only 43.5 s on 8,192 cores, achieving a speedup of 2,368-fold. If we increase the number of characters by a factor of 10, ExaBayes scales from 256 up until at least 32,768 cores and improves runtime even faster than theoretically expected for up to 4,096 cores because of increased cache efficiency (see fig. 2).

ExaBayes implements Metropolis-coupling (Geyer 1992), a fundamental mechanism to accelerate convergence on "difficult" data sets. Heated chains with increased acceptance probabilities are coupled to the chain that is being sampled. Swaps between chains are accepted proportional to the posterior probability of chain states. As memory requirements increase linearly with each additional chain, large data sets require chains to be distributed across several CPUs or computing nodes in a cluster (second layer of parallelism). To this end, we modified the state-of-the-art parallel algorithm for Metropolis-coupled MCMC (Altekar et al. 2004) to use faster, nonblocking communication among processes (see supplementary material, Supplementary Material online, for a detailed description). Thus, when a set of processors propose a swap through a nonblocking message, they can immediately continue doing useful work by conducting calculations on a different chain that will not be swapped. This reduces runtimes by up to 19% and thus avoids wasting hundreds of CPU hours for large-scale analyses. Finally, independent analyses can be executed concurrently using a third layer of parallelism (see fig. 1).

Moreover, ExaBayes adapts two orthogonal memory saving techniques. The so-called subtree equality vector approach allows to save memory proportional to the fraction of unknown or missing data. At the same time, it can also accelerate likelihood calculations on data sets with a large proportion of missing data. This strategy was designed for phylogenomic data sets (Izquierdo-Carrasco et al. 2011), where unknown orthologs can lead to a large fraction of missing data exceeding for instance 75% (DellAmpio et al. 2014). Our second technique allows discarding memory-intensive partial results by recomputing these on demand. Three settings trade varying amounts of runtime for additional memory savings by means of recomputation. Specifically, when coupled chains are distributed, this reduces memory requirements by a factor of up to 2, while we observe a slow-down of less than 1.5-fold. As the size of data sets that can be computed by ExaBayes is merely constrained by available hardware, our memory saving techniques allow conducting ambitious analyses on small and less expensive computer clusters.

For improved model fit, it is often desirable to sample distinct model parameters (e.g., substitution rates) for each gene or partition of an alignment. We have tested ExaBayes using simulated alignments with 1,000–10,000 partitions. All parameters (including branch lengths, but excluding the topology) can be flexibly linked or unlinked across partitions. We observed a performance decrease with an increasing number of partitions. To alleviate this issue, we modified our MCMC algorithm and data-to-processor assignment scheme. This modification induces a runtime improvement of up to 22 times for runs in which branch lengths are linked across all partitions. ExaBayes runs up to 87 times faster, when each partition has distinct branch lengths (i.e., branch lengths are unlinked across partitions).

Inference from a Whole-Genome Data set

To demonstrate the capabilities of ExaBayes, we executed a Bayesian inference on a "difficult" complete simulated

genome (for which the "true" evolutionary history is known) comprising 100 partitions with 1,000,000 characters each (inspired by a per-chromosome partitioning). To simulate the alignment, we used a tree with 200 species. The tree is bush-like and contains many long, outer branches and short, hard-to-resolve, inner branches. As the RAM requirements of this data set exceed 5 TB, we employed the SuperMUC supercomputer, which is currently among the ten fastest supercomputers in the world. A total of four independent runs with one chain each seeded by reasonable (i.e., nonrandom) parsimony starting trees (Fitch and Margoliash 1967) rapidly converged to the true tree topology within less than 20,000 generations (total chain length: 100,000 generations; a sample was extracted every 500 generations). All branches showed 100% certainty (posterior probability). Using over 4,096 CPU cores, the slowest run took 1 h 40 min. Thus, the accumulated CPU-time over all four runs is 3 years and 45 days. To avoid potential biases induced by the parsimony starting trees, we also ran two independent chains starting from random trees. Here, the chains converged to the true tree topology after $\approx$60,000 generations. With random starting trees, we have to discard substantially longer burn-ins, before we attain an accurate sampling of the posterior probability of the trees.

We examined how the posterior probability of the trees changes as we reduce the amount of data in steps of 2 orders of magnitude. If we discard 33% of samples as burn-in, we obtain 1) 100% confidence for all splits in the tree inferred from 100,000,000 characters, 2) only one split with 98.5% certainty for 1,000,000 characters, and 3) nine splits with a posterior probability between 61.2% and 99.2% for 10,000 characters. For 10,000 characters we do not attain maximum confidence in the true tree, even if the burn-in phase is substantially extended. Thus, we conclude that this amount of data is insufficient for inferring a reliable tree on this data set. As illustrated in figure 3, the chain instead jumps between several trees of high posterior probability after burn-in.

Finally, we examined how confidence intervals for branch lengths change as we increase the amount of data. We find that, with only 10,000 characters, we mostly do not obtain accurate branch length estimates and that confidence intervals cover a wide spectrum specifically for short branch lengths. Branch lengths can be determined more accurately with more data; however, increasing the data set size from 1,000,000 characters to 100,000,000 characters does not substantially decrease the standard deviation in branch length samples (see supplementary material, Supplementary Material online). We observed that genome-scale data allow for high certainty about the "true" (here simulated) branch length; however, even this amount of data does not guarantee small confidence intervals for branch lengths in the range of $10^{-3} - 10^{-4}$. Nevertheless, we expect that with chromosome- and genome-size data, extremely precise estimations of divergence times can be performed.

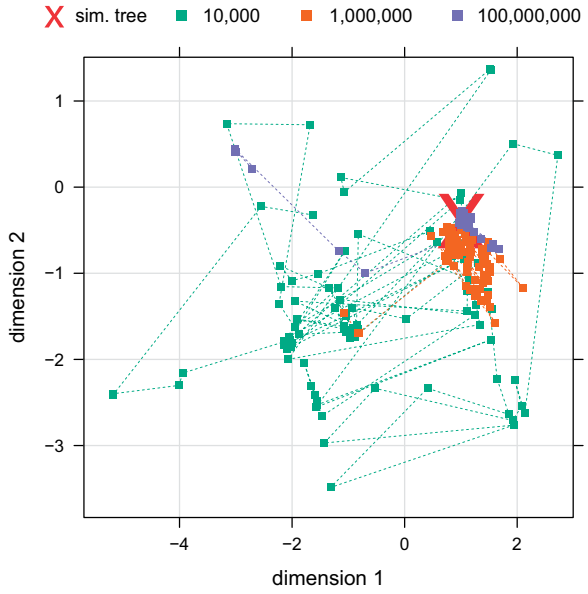

FIG. 3. Inference from genome-sized data: 2D-rescaled representation employing multidimensional scaling (MDS) of Robinson–Foulds distances among sampled trees for chains with varying amount of data. The position of the simulated true tree is shown in red. The applied MDS algorithm maps identical trees to adjacent nonidentical positions (i.e., overlapping squares represent identical trees).

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors are extremely grateful to Fredrik Ronquist for sharing his knowledge and expertise about Bayesian phylogenetic analysis with them. They thank Paschalia Kapli for β-testing the program and helping to improve the user-friendliness of `ExaBayes`. Experiments for demonstrating scalability and the whole-genome inference were covered by a CPU-time grant for the SuperMUC system at the Leibniz Supercomputing Centre (LRZ). All authors are funded through institutional funding provided by the Heidelberg Institute for Theoretical Studies.

References

Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F. 2004. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* 20(3):407–415.

DellAmpio E, Meusemann K, Szucsich NU, Peters RS, Meyer B, Borner J, Petersen M, Aberer AJ, Stamatakis A, Walzl MG, et al. 2014. Decisive data sets in phylogenomics: lessons from studies on the phylogenetic relationships of primarily wingless insects. *Mol Biol Evol.* 31(1):239–249.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol.* 29(8):1969–1973.

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, et al. 2008. Broad

phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452(7188):745–749.
Fitch WM, Margoliash E. 1967. Construction of phylogenetic trees. *Science* 155(760):279–284.
Geyer CJ. 1991. Markov chain Monte Carlo maximum likelihood. Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface. p. 156–163.
Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. *Syst. Biol.* 59(3):307–321.
Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57(1):97–109.
Izquierdo-Carrasco F, Smith SA, Stamatakis A. 2011. Algorithms, data structures, and numerics for likelihood-based phylogenetic inference of huge trees. *BMC Bioinformatics* 12(1):470.
Lakner C, van der Mark P, Huelsenbeck JP, Larget B, Ronquist F. 2008. Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. *Syst. Biol.* 57(1):86–103.
Rambaut A. 2014. Figtree, a graphical viewer of phylogenetic trees [Internet]. Available from: http://tree.bio.ed.ac.uk/software/figtree.
Rambaut A, Drummond A. 2007. Tracer v1.4 [Internet]. Available from http://tree.bio.ed.ac.uk/software/tracer/.
Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, et al. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499:431–437.
Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61(3):539–542.
Stamatakis A. 2014. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect Math Life Sci.* 17:57–86.
Yang Z. 1994a. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol.* 39(3):306–314.
Yang Z. 1994b. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Syst Biol.* 43(3):329–342.